Inter-observer agreement in the evaluation of digitized cervical images.

Jose Jeronimo, M.D.[1¶], L. Stewart Massad, M.D.[2], Philip E. Castle, PhD, MPH[1], Mark Schiffman, M.D.[1] for the National Institutes of Health / American Society for Colposcopy and Cervical Pathology (NIH/ASCCP) Research Group[*].

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.
[2] Dept. of Obstetrics & Gynecology, Southern Illinois University School of Medicine, Springfield, IL.

Abstract:

Background: Observational agreement between clinicians has been evaluated in several medical specialties with very different levels of concordance. Previous reports have shown lack of agreement among colposcopists.

Methods: Twenty expert colposcopists evaluated 939 digitized images of the uterine cervix obtained after the application of 5% acetic acid during the ASCUS-LSIL Triage Study. Pictures were distributed among the evaluators such that each expert had images with similar visual diagnoses taken from women with similar HPV DNA test results. Each evaluated 112 pictures. Twenty images were graded by all the colposcopists. McNemar's chi square tests were used to compare ratings of lesion presence, lesion number, and diagnosis. Factors associated with diagnosis were determined using logistic regression.

Results: Frequency of diagnoses varies widely among evaluators ($p<0.001$) with the greatest variability for normal/metaplasia diagnosis. Despite the wide variability of

diagnoses, it is possible to create up to three groups of evaluators who tended to resemble each other in their diagnosis.

Discussion: We conclude that colposcopic diagnosis using static images is irreproducible and might reflect similar problems in clinical practice. We recommend that researchers question the use of colposcopic images as a reference standard for teaching and evaluating the presence or severity of disease.